OPEN ACCESS

University of Dundee

University of Dundee

Cross-scenario transfer person reidentification

Wang, Xiaojun; Zheng, Wei-Shi; Li, Xiang; Zhang, Jianguo

Published in:
IEEE Transactions on Circuits and Systems for Video Technology

DOI:
10.1109/TCSVT.2015.2450331

Publication date:
2016

Document Version
Peer reviewed version

Link to publication in Discovery Research Portal

Citation for published version (APA):
Wang, X., Zheng, W-S., Li, X., & Zhang, J. (2016). Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*(8), 1447-1460. https://doi.org/10.1109/TCSVT.2015.2450331

General rights
Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Cross-scenario Transfer Person Re-identification

Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang

Abstract—Person re-identification is to match images of the same person captured in disjoint camera views and at different time. In order to obtain a reliable similarity measurement between images, manually annotating a large amount of pairwise cross-camera-view person images is deemed necessary. However, such a kind of annotation is both costly and impractical for efficiently deploying a re-identification system to a completely new scenario, a new setting of non-overlapping camera views between which person images are to be matched. To solve this problem, we consider utilizing other existing person images captured in other scenarios to help the re-identification system in a target (new) scenario, provided that a few samples are captured under the new scenario. More specifically, we tackle this problem by jointly learning the similarity measurements for re-identification in different scenarios in an asymmetric way. To model the joint learning, we consider that the re-identification models share certain component across tasks. A distinct consideration in our multi-task modeling is to extract the discriminant shared component that reduces the cross-task data overlap in the shared latent space during the joint learning, so as to enhance the target inter-class separation in the shared latent space. For this purpose, we propose to maximize the cross-task data discrepancy (CTDD) on the shared component during asymmetric multi-task learning, along with maximizing the local inter-class variation and minimizing local intra-class variation on all tasks. We call our proposed method the constrained asymmetric multi-task discriminant component analysis (cAMT-DCA). We show that cAMT-DCA can be solved by a simple eigen-decomposition with a closed form, getting rid of any iterative learning used in most conventional multi-task learning. The experimental results show that the proposed transfer model gains a clear improvement against the related non-transfer and general multi-task person re-identification models.

Index Terms—person re-identification, cross-scenario transfer, visual surveillance

1 Introduction

Surveillance systems have become almost ubiquitous in large public spaces, even in private places [1], [2]. A most recent topic generating more and more interests in surveillance is person re-identification (Re-ID). Re-ID aims to re-identify an individual who has been previously observed over spatially disjoint cameras views in a wide area surveillance system, which is an important task for continuous object tracking and human behavior analysis over large-scale camera networks. Notable progress has been made in Re-ID in the last few years, including the attempt of designing cross-view invariant and discriminant features [3]–[16] and developing metric learning methods for the similarity measurement between images across non-overlapping camera views [17]–[27].

One of the essential requirements for a Re-ID system is its deployability and applicability in a new scenario[1] (e.g., from the street to the underground). However, this goal has hardly been achieved. The main challenges are different *lighting* conditions, the change of camera viewing *angles*, *posture* variation and *occlusion* change, etc. The raw features shown robust for one type of scene do not always work well for another type of scene. The model of a Re-ID system often needs to be built with capability of feature learning either explicitly or implicitly. Furthermore, to train a robust Re-ID system, one consensus is to collect a large amount of labeled training data, ideally for the specific working scenario considered. Though this could work well in principle, generating such a dataset at a large scale in Re-ID is not a trivial task, as one needs to manually label pairwise pedestrian images from different camera views, and this often involves another difficult task, which is tracking by hand each person across non-overlapping camera views. Therefore, it is cost prohibitive, especially in a crowded public space such as airport. There exist many datasets already collected for training Re-ID systems in different scenarios. Although each of them might be of its own limited scope, a question arises: can they help enrich each other? In other words, is it possible to use other datasets collected at different scenarios to enrich the learning on a target one? In addition, considering utilizing other datasets to assist the person re-identification on a target one is also helpful for deploying a new Re-ID system in a new scenario shortly without heavy annotation for this new scenario. In this work, we call the problem of transferring data collected from other scenarios to help set up a Re-ID system in a new scenario without re-collecting a lot of labeled data as the *cross-scenario transfer person re-identification*.

If we consider training a Re-ID system with data from a new scenario as a *target task* and the training on an existing source dataset from another scenario elsewhere as a *source task*, then using the learning on large amounts of source data to improve the target one can be treated as an asymmetric multi-task learning (MTL) problem [28], where "asymmetric" means the joint learning does not aim to benefit both target and source tasks but mainly the target one. The underlying motivation is that, we consider pedestrians in all scenarios

- *X. Wang, W.-S. Zheng and X. Li are with the School of Information Science and Technology, Sun Yat-Sen University, Guangzhou, China. E-mail: xiaojuanwang.cs@gmail.com, wszheng@ieee.org and lixiang651@gmail.com*
- *J. Zhang is with the School of Computing, University of Dundee, UK. E-mail: j.n.zhang@dundee.ac.uk*

1. In this work, a *scenario* is regarded as a setting of non-overlapping camera views from which person images are to be matched.

share certain common variations such as pose variations which are potentially independent of the scenarios. And the shared variations could be preserved in a low rank dimensional subspace, which we call the shared latent space. As typical multi-task learning does, there is a common component shared by all tasks and there is an individual specific component for each task to constitute the similarity measurement. In this work, we consider learning the common component used for both tasks by exploring a shared latent subspace and consider learning each task-specific component by exploring a task-specific subspace.

Different from existing multi-task learning methods, we particularly consider further exploring discriminant modeling in the shared latent space. That is to say, we consider separation of different task data points in the shared latent space for our cross-scenario transfer person re-identification modeling. This is motivated by the fact that sometimes the intra-class variation and inter-class variation are similar [25]. As shown in our analysis (see Fig. 2 and Fig. 3 for details), data samples from different tasks could overlap in the shared latent space. This is not a desired behavior as the person identities from different tasks are usually different. Unfortunately, such a problem is not investigated in existing multi-task learning methods. To solve this problem, we propose a *cross-task data discrepancy* (CTDD) criterion to measure the discrepancy across tasks in the shared latent subspace, so as to enhance the target inter-class separation modeling there.

With the above ideas, we learn the similarity measurement for each task in cross-scenario transfer person re-identification by a joint learning of the shared latent subspace and the corresponding task-specific subspace. Since the appearances of person images have dramatic variations due to changes caused by pose, action, and lighting, we hope the learned measurement is locally discriminant. Therefore, we maximize local inter-class variation and CTDD, and meanwhile minimize local intra-class variation. We call our method the *constrained asymmetric multi-task discriminant component analysis* (cAMT-DCA) model. Distinct to most existing multi-task learning methods that optimize the objective functions through an iterative technique, we present the derivation of a closed-form solution and thus a globally optimal solution can be guaranteed. An illustration of our proposed asymmetric multi-task Re-ID system is shown in Fig. 1.

We evaluate the proposed model on different cross-scenario transfer settings combined by commonly used Re-ID datasets, including 3DPeS [29], i-LIDS [25], CAVIAR [30] and VIPeR [8]. The results on Re-ID demonstrate that the proposed method is better than other related state-of-the-art methods.

2 Related Work

Person re-identification has received considerable attention in recent years. Given a *query* image (also known as a *probe* image), the task is to find its best match from a pool of candidate images (also known as *gallery* images) captured from different camera views. Most of the research to date follows a two-step paradigm. Firstly, a feature representation is built for each image, and the query is paired with each of the gallery images. Secondly, the similarity of images in each pair is calculated based on a certain metric, which is then used as a ranking criterion to determine whether a gallery image contains the same person as the query image. The majority of existing methods focus on either building invariant and robust feature representations or developing reliable metrics for matching [25], [31]. In the following, we review previous works that are most relevant to ours.

Metric learning methods in person re-identification are related to ours in the sense of learning a robust metric to obtain a reliable similarity measurement. Among existing works, some classical metric learning methods in machine learning were either adopted or further developed for Re-ID. Dikmen et al. [20] introduced a Large Margin Nearest Neighbor method, with a rejection option (LMNN-R) to directly learn the Mahalanobis metric by extending LMNN [17]. Information Theoretic Metric Learning (ITML) [18] could also be applied in this framework. Kostinger et al. [22] proposed a novel method KISSME to learn a distance metric from equivalence constraints from a statistical inference perspective, and later a regularized KISS metric learning was further developed [24]. Mignon et al. introduced Pairwise Constraint Component Analysis (PCCA) [21], which learns a projection into a low-dimensional space to deal with the high-dimensional input space. Li et al. [27] proposed to learn locally-adaptive decision functions (LADF) for person verification that can be viewed as a joint model of distance metric and a locally adaptive thresholding rule. The relative-comparison based methods [19], [25] in person re-identification have also been proposed. In particular, Zheng et al. [25] formulated person re-identification as a relative distance comparison (RDC) learning problem. RDC is formulated to maximize the likelihood of a pair of true matches having a relatively smaller distance than a pair of wrong matches in a soft discriminant manner. Nonetheless, the objective function is not convex and thus a global optimal solution is not guaranteed. Furthermore, the computational cost is very high because image distances of every pair have to be compared. Recently, a subspace learning method closely related to distance learning, Local Fisher Discriminant Analysis (LFDA) [26], [32], has been applied to person re-identification and encouraging results on a few datasets were reported. All these methods mentioned above focused on the problem of Re-ID in a single scenario and do not specially address the challenge of cross-scenario transfer person re-identification.

Relevant to our multi-task formulation, there are some existing multi-task metric learning methods either for Re-ID such as multi-task maximally collapsing metric learning (MtMCML) [31] or for general purpose such as MT-LMNN [33] and GPLMNN [34], but they were not developed for transfer across Re-ID datasets. In particular, MtMCML aimed to overcome an existing disadvantage of using metric in Re-ID where only one unique metric was assigned to all image pairs without considering the different settings of cameras. To solve this, MtMCML designed one metric for each camera pair and learned a set of multiple metrics jointly. However, one of its prerequisites is that different tasks should share the same label set, which is obviously not applicable to the

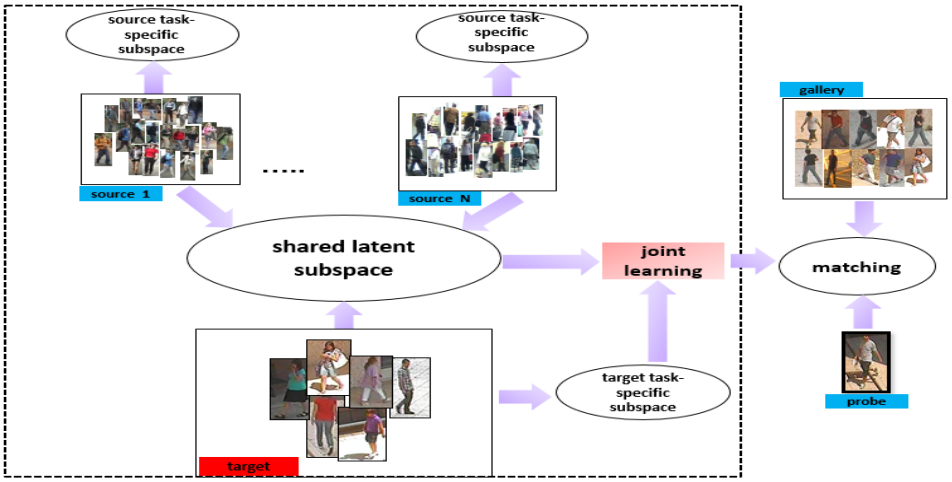


Fig. 1: Asymmetric Multi-task Person Re-identification system: diagrams inside the large dash rectangle box indicate the principle of training.

problem of cross-scenario transfer. For those multi-task metric learning models in machine learning community, firstly, they are not specifically designed for person re-identification, thus with unknown applicability. Secondly, these multi-task metric models differ from ours in that although they also learn a common component across tasks, they do not particularly ensure that the common component itself is discriminant but only care that the combination of the task-specific component and common component as a whole is discriminant. In our work, we show that reducing the overlap of data across tasks in the latent shared subspace further enhances the transfer from auxiliary data, resulting in a notably better identification performance in our experiments.

It is noted that there are other related works on attempting transfer learning in Re-ID. Zheng et al. [35] considered using other non-target data to improve the Re-ID performance on target data. However, their modeling is only for watch-list based transfer where the *gallery* images of probe people are required during training, while our focus is on cross-scenario transfer where any *probe* information is assumed unknown during training. Li et al. [36] proposed to learn candidate-specific metrics for inter-camera-view matching and addressed this by transferring the metrics learned from subsets of training images that are visually similar to the candidates. However, this technique assumes all candidates are from the same scenario as the training data, and does not seek to resolve the source-target difference issue. Therefore, the transferred metric may not be generalizable across scenarios. There exist quite a few other related subspace/metric learning based transfer learning methods, such as TCA [37], TFLDA [38] and DAML [39]. However, these methods are not for transfer of general purpose but particularly designed for domain adaptation which assumes that samples from the target domain (i.e. target dataset) and source domains (i.e. datasets collected in other scenarios) share the same label sets. In principle, the cross-scenario transfer person re-identification is not a domain adaptation problem, since persons present in the target task do not necessarily appear in the source tasks in most cases. And thus the underlying assumption of data label sharing across tasks made by domain adaptation methods is not applicable.

In summary, our main contributions are:

1) We formulate the problem of cross-scenario transfer person Re-ID as a model of asymmetric multi-task learning. To the best of our knowledge, it is the first attempt to address this issue, since it is a largely unaddressed problem in person re-identification.
2) A constrained asymmetric multi-task discriminant component analysis model (cAMT-DCA) has been proposed, in which the Cross-task Data Discrepancy (CTDD) is designed for learning a discriminant shared latent space.
3) We show that our multi-task formulation can be solved by a generalized eigen-decomposition, so that a globally optimal solution can be obtained.

3 Cross-scenario Transfer Modeling for Person Re-identification

Here we refer the datasets collected elsewhere as source datasets and the datasets collected in our concerned surveillance system as target datasets. Without loss of generality, we firstly elucidate our method in the case where only one source is available. We then present the formulation in the case of multiple sources.

Notations. In the reminder of this paper, we use the superscript $'$ to denote the transpose of a vector or a matrix. We define $\mathbf{I}_d$ as the $d \times d$ identity matrix and $\mathbf{O}_{d\times m}$ as the $d \times m$ matrix of all zeros.

3.1 Transfer One Source Dataset

Let $\mathcal{X}_s$ and $\mathcal{X}_t$ be the labeled source dataset and target dataset, respectively. Samples in both datasets are of dimensionality d. Let n_s and n_t be the numbers of samples in $\mathcal{X}_s$ and $\mathcal{X}_t$ respectively, with $n_s \gg n_t$.

The purpose of cross-scenario transfer Re-ID is to train a robust system on a limited target dataset by leveraging existing relevant source datasets, which are often captured in

different scenarios. This requirement could be assembled well by the prime idea of asymmetric MTL [28], which aims to train multiple related tasks simultaneously and mainly benefit the target one from the propagation of the shared information across tasks. Motivated by this, we propose to model the commonalities between the target and source datasets through a shared latent subspace, spanned by the columns of a projection matrix $\mathbf{W}_0 \in \mathrm{R}^{d\times r}$. We also introduce a task-specific subspace through a task-specific projection matrix $\mathbf{W}_s$ for the source dataset and a projection matrix $\mathbf{W}_t$ for the target dataset. $\mathbf{W}_s$ and $\mathbf{W}_t$ are set to have the same size as $\mathbf{W}_0$ to ensure that the resulting subspaces have the same dimensionality, and the projected features can be easily compared. Intuitively, the shared latent space defined by $\mathbf{W}_0$ is a critical component beneficial to the learning on limited target data during the transfer. In this way, the projection of a target sample $\mathbf{x}_t$ could be represented as a linear combination of the projections in the two subspaces by

$$\mathbf{z}_t = ((1-\beta)\mathbf{W}_0 + \beta\mathbf{W}_t)'\mathbf{x}_t, \tag{1}$$

where $0 \leq \beta \leq 1$ is to explicitly control the strength of the connection between the shared latent projection and the target data-specific projection during the unification. Similarly, a source sample $\mathbf{x}_s$ is represented by the projections in the subspaces as

$$\mathbf{z}_s = ((1-\beta)\mathbf{W}_0 + \beta\mathbf{W}_s)'\mathbf{x}_s, \tag{2}$$

With the above formulation, we wish to perform a joint learning of the projections in Eq. (1) and Eq. (2) such that the intra-class variance is minimized and the inter-class variance is maximized simultaneously in both tasks after the projections. We start the formulation from modeling local intra- and inter-class variations. More specifically, as in [26], [40], let $\mathbf{S}_b^s$ and $\mathbf{S}_w^s$ denote the local inter-class and intra-class covariance matrices on source dataset, respectively, and let $\mathbf{S}_b^t$ and $\mathbf{S}_w^t$ be the ones on target dataset. Those locally-weighted scatter matrices are computed as follows:

$$\mathbf{S}_b^q = \frac{1}{2}\sum_{i,j=1}^{n} \overline{\mathbf{A}}_{i,j}^b (\mathbf{x}_i^q - \mathbf{x}_j^q)(\mathbf{x}_i^q - \mathbf{x}_j^q)' \tag{3a}$$

$$\mathbf{S}_w^q = \frac{1}{2}\sum_{i,j=1}^{n} \overline{\mathbf{A}}_{i,j}^w (\mathbf{x}_i^q - \mathbf{x}_j^q)(\mathbf{x}_i^q - \mathbf{x}_j^q)' \tag{3b}$$

where $q \in \{s,t\}$ is the task indicator. Let n_c be the number of samples of class c and n the total number of samples from all classes. In the above formulation, each pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is weighted based on their affinity $\mathbf{A}_{i,j}$, which could be computed as follows: $\overline{\mathbf{A}}_{i,j}^b = \frac{\mathbf{A}_{i,j}}{n} - \frac{\mathbf{A}_{i,j}}{n_c}$ and $\overline{\mathbf{A}}_{i,j}^w = \frac{\mathbf{A}_{i,j}}{n_c}$ if $\mathbf{x}_i, \mathbf{x}_j$ are from the same class, and $\overline{\mathbf{A}}_{i,j}^b = \frac{1}{n}$ and $\overline{\mathbf{A}}_{i,j}^w = 0$ if otherwise.

In order to jointly maximize the ratio between inter-class covariance and intra-class covariance in *both* target and source datasets, we propose the following objective function:

$$\max_{\mathbf{W}_1,\mathbf{W}_2} (1-\gamma)\frac{\mathrm{tr}(\mathbf{W}_1'\mathbf{S}_b^s\mathbf{W}_1)}{\mathrm{tr}(\mathbf{W}_1'\mathbf{S}_w^s\mathbf{W}_1)} + \gamma\frac{\mathrm{tr}(\mathbf{W}_2'\mathbf{S}_b^t\mathbf{W}_2)}{\mathrm{tr}(\mathbf{W}_2'\mathbf{S}_w^t\mathbf{W}_2)}, \tag{4}$$

where $\mathbf{W}_1 = (1-\beta)\mathbf{W}_0 + \beta\mathbf{W}_s, \mathbf{W}_2 = (1-\beta)\mathbf{W}_0 + \beta\mathbf{W}_t$. The first term measures the separability (Fisher criterion) on the source dataset, while the second term measures the separability on the target dataset. Note that here the weighted average through γ is actually on the score level rather than on the feature-level because of the trace operation. It controls the contribution of source and target data in the objective function. The transfer is mainly achieved through the shared latent projection component $\mathbf{W}_0$ in Eq. (1) and Eq. (2). This shared component and the task-specific components are simultaneously learned through the above asymmetric joint optimization parameterized by γ.

Based on the formula (4), we can have a further insight into our multi-task modeling in Eq. (1) and Eq. (2). Taking the source inter-class variance $\mathrm{tr}(\mathbf{W}_1'\mathbf{S}_b^s\mathbf{W}_1)$ as an example (intra-class variance could be illustrated in the same way), $\mathrm{tr}(\mathbf{W}_1'\mathbf{S}_b^s\mathbf{W}_1)$ can be rewritten as

$$\begin{aligned}
&\mathrm{tr}(\mathbf{W}_1'\mathbf{S}_b^s\mathbf{W}_1) \\
&= \frac{1}{2}\sum_{i,j=1}^{n} \overline{\mathbf{A}}_{i,j}^b \sum_{k=1}^{r} \mathbf{W}_1(:,k)'(\mathbf{x}_i^s - \mathbf{x}_j^s)(\mathbf{x}_i^s - \mathbf{x}_j^s)'\mathbf{W}_1(:,k) \\
&= \frac{1}{2}\sum_{i,j=1}^{n} \overline{\mathbf{A}}_{i,j}^b \sum_{k=1}^{r} \left[(1-\beta)\mathbf{W}_0(:,k)'(\mathbf{x}_i^s - \mathbf{x}_j^s) + \beta\mathbf{W}_s(:,k)'(\mathbf{x}_i^s - \mathbf{x}_j^s)\right]^2
\end{aligned}$$

where $\mathbf{W}_1(:,k)$ is the k^{th} column of $\mathbf{W}_1$. From the above we can see the projection scores of each sample difference (e.g. $\mathbf{x}_i^s - \mathbf{x}_j^s$) onto each of $\mathbf{W}_0$ and each of $\mathbf{W}_s$ (or $\mathbf{W}_t$) are actually measured on the discriminative directions (shared and task-specific), and adding those measures together gives us a stronger cue on overall discriminativeness.

Although the proposed multi-task method differs from existing ones, the weighting strategy is often used in other multi-task learning methods for general purpose such as MT-LMNN [33]. If $\gamma > 0.5$, target task is treated more importantly than the source task in the objective function. Therefore, more learning efforts are put on the target task in optimization. Once the optimal $\mathbf{W}_2$ is learned, we can perform re-identification on the target dataset after the projection in Eq. (1) using the simple Euclidean distance between projected features.

However, the above objective function is non-convex, and it is difficult to find a globally optimal solution, since $\mathbf{W}_1$ and $\mathbf{W}_2$ are not independent and are shared by a common component $\mathbf{W}_0$. To make it tractable, we propose a relaxed objective function as follows:

$$\max_{\mathbf{W}_1,\mathbf{W}_2} \frac{\mathrm{tr}((1-\gamma)\mathbf{W}_1'\mathbf{S}_b^s\mathbf{W}_1 + \gamma\mathbf{W}_2'\mathbf{S}_b^t\mathbf{W}_2)}{\mathrm{tr}((1-\gamma)\mathbf{W}_1'\mathbf{S}_w^s\mathbf{W}_1 + \gamma\mathbf{W}_2'\mathbf{S}_w^t\mathbf{W}_2)}, \tag{5}$$

Compared to formula (4), the above formulation maximizes the joint inter-class covariances of source and target tasks and meanwhile minimizes their joint intra-class covariances. Hence to some extent it also reflects the separability of the projected source data and target data. A trade-off parameter $0 \leq \gamma \leq 1$ is also used to control the learning strength of each task. If $\gamma = 1$, the learning is only performed on target dataset, and if $\gamma = 0$, it is only on source dataset. Hence, we call the above model as *asymmetric multi-task discriminant component analysis* (AMT-DCA).

Note that when $\beta = 1, \gamma = 1$, the model degrades to LFDA [26], [32] on the target dataset in a single task setting.

However, LFDA is not a multi-task method and its original formulation cannot cope with the multi-task learning.

Optimization. How to optimize the objective function in formula Eq. (5) with respect to $\mathbf{W}_0, \mathbf{W}_s$ and $\mathbf{W}_t$ simultaneously seems a challenging problem. However, by using the form of block matrix, we show that the above problem can be solved by a simple eigen-decomposition. In particular, let $\mathbf{W} = [\mathbf{W}_0; \mathbf{W}_s; \mathbf{W}_t] \in \mathbb{R}^{3d\times r}$, and $\mathbf{\Theta}_s = [(1-\beta)\mathbf{I}_d, \beta\mathbf{I}_d, \mathbf{O}_{d\times d}] \in \mathbb{R}^{d\times 3d}$, $\mathbf{\Theta}_t = [(1-\beta)\mathbf{I}_d, \mathbf{O}_{d\times d}, \beta\mathbf{I}_d] \in \mathbb{R}^{d\times 3d}$, the optimization problem in Formula (5) can be converted to

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}'\mathbf{A}\mathbf{W})}{\mathrm{tr}(\mathbf{W}'\mathbf{B}\mathbf{W})} \tag{6}$$

where

$$\mathbf{A} = (1-\gamma)(\mathbf{\Theta}_s^{'}\mathbf{S}_b^s\mathbf{\Theta}_s) + \gamma(\mathbf{\Theta}_t^{'}\mathbf{S}_b^t\mathbf{\Theta}_t) \tag{7a}$$

$$\mathbf{B} = (1-\gamma)(\mathbf{\Theta}_s^{'}\mathbf{S}_w^s\mathbf{\Theta}_s) + \gamma(\mathbf{\Theta}_t^{'}\mathbf{S}_w^t\mathbf{\Theta}_t) \tag{7b}$$

Both $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite matrices, making the globally optimal solution guaranteed. Formula (6) is a well-known form and could be solved by a typical generalized eigenvalue problem:

$$\mathbf{A}\mathbf{W} = \lambda\mathbf{B}\mathbf{W} \tag{8}$$

Compared to other multi-task learning methods, a very prominent property of our method is that we have a closed-form solution. In practice, if $\mathbf{B}$ becomes non-invertible, a simple perturbation can be done to avert it.

3.2 Transfer Multiple Source Datasets

Suppose that there are m source datasets available, and $\mathcal{X}_s^i$ is the i-th labeled source dataset. Let $\mathbf{S}_b^{s,i}$ and $\mathbf{S}_w^{s,i}$ be the corresponding local inter-class and intra-class covariance matrices of the i-th labeled source dataset, respectively, $i = 1, 2, \ldots, m$. Similarly, we introduce the task-specific projection $\mathbf{W}_s^i$ for each source dataset and $\mathbf{W}_t$ for the target dataset. All tasks are connected by a shared latent projection $\mathbf{W}_0$. Similarly, we define

$$\mathbf{W} = [\mathbf{W}_0; \mathbf{W}_s^1; \ldots; \mathbf{W}_s^i; \ldots; \mathbf{W}_s^m; \mathbf{W}_t] \in \mathbb{R}^{(m+2)d\times r}. \tag{9}$$

Let $\mathbf{\Theta}_s^i = [(1-\beta)\mathbf{I}_d, \ldots, \beta\mathbf{I}_d, \ldots, \mathbf{O}_{d\times d}] \in \mathbb{R}^{d\times(m+2)d}$, and $\mathbf{\Theta}_s^i$ can be partitioned into $1\times(m+2)$ sub-matrices of size $d\times d$, where the first sub-matrix is $(1-\beta)\mathbf{I}_d$, the last is a zero matrix $\mathbf{O}_{d\times d}$, and all of the sub-matrices in the between are $\mathbf{O}_{d\times d}$ except for the i_{th} sub-matrix, which is $\beta\mathbf{I}_d$. Here $\mathbf{\Theta}_t = [(1-\beta)\mathbf{I}_d, \mathbf{O}_{d\times d}, \ldots, \mathbf{O}_{d\times d}, \beta\mathbf{I}_d]$. To perform a joint learning on target and multiple source datasets, we derive an objective function of the same form as in Eq. (6) by redefining

$$\mathbf{A} = (1-\gamma)(\frac{1}{m}\sum_{i=1}^{m}(\mathbf{\Theta}_s^i)'\mathbf{S}_b^{s,i}\mathbf{\Theta}_s^i) + \gamma(\mathbf{\Theta}_t^{'}\mathbf{S}_b^t\mathbf{\Theta}_t) \tag{10a}$$

$$\mathbf{B} = (1-\gamma)(\frac{1}{m}\sum_{i=1}^{m}(\mathbf{\Theta}_s^i)'\mathbf{S}_w^{s,i}\mathbf{\Theta}_s^i) + \gamma(\mathbf{\Theta}_t^{'}\mathbf{S}_w^t\mathbf{\Theta}_t). \tag{10b}$$

Similarly, the solution can be obtained by Eq. (8).

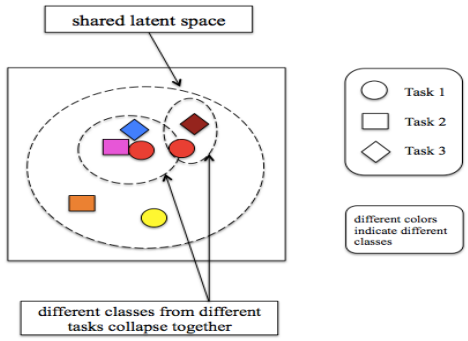


Fig. 2: A schematic illustration of the motivation behind CTDD. Different shapes represent data from different tasks, and different colors represent different classes. In the shared latent space, different classes from different tasks could collapse together.

4 Constrained Asymmetric Multi-task Discriminant Component Analysis

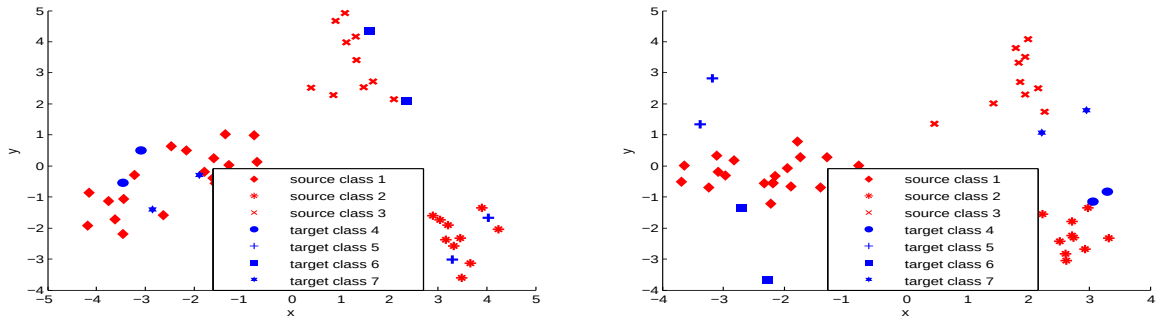
A key idea of MTL is to share common component across tasks. Particularly in multi-task metric learning, the common component induces a shared low-dimensional representation across tasks, which means data points across tasks are all projected to the same subspace, namely the shared latent subspace. The shared latent subspace can be used for all intra-task discriminant modeling. However, none of the existing methods has ever considered separating data between different tasks in the shared latent subspace. As a result, data samples from different tasks may collapse together in that shared latent space (see Fig. 2 and Fig. 3). This is not a desired behavior in cross-scenario transfer Re-ID. We are interested in separating the data from different tasks as they normally represent different cohorts of people in different scenarios. Therefore, intuitively, maximizing the difference between samples of different tasks can enhance the discriminant modeling on target data in the shared latent space.

To solve this problem, we further propose a discriminant model for separating data of different tasks. Specifically, we consider the cross-task data separation by introducing a criterion called *cross-task data discrepancy* (CTDD) in the shared latent subspace induced by the columns of $\mathbf{W}_0$ as below:

$$CTDD(\mathbf{W}_0) = \frac{1}{N}\,\mathrm{tr}(\mathbf{W}_0^{'}\{\sum_{k\neq l}\sum_{i,j}(\mathbf{x}_i^k - \mathbf{x}_j^l)(\mathbf{x}_i^k - \mathbf{x}_j^l)'\}\mathbf{W}_0) \tag{11}$$

where k and l are *task indices*, i, j are indices of samples in each task (e.g. $\mathbf{x}_i^k$ denotes the i_{th} sample in the task k), and N is the total number of cross-task image pairs.

We illustrate our motivation in Fig. 3 by taking the transfer from CAVIAR to i-LIDS as an example. When there is no CTDD, namely no cross-task data separation imposed in the shared latent space, we got $\mathbf{W}_0$ based on Eq. (5), then randomly selected a set of persons in source dataset (in red colour) and in target dataset (in blue colour), and finally projected them into the shared latent space $\mathbf{W}_0$. The 2D visualization of these samples in the shared latent space is achieved using PCA. As shown in Fig. 3(a), some target



(a) Without CTDD (b) With CTDD

Fig. 3: Illustration of the effect of CTDD in the transfer from CAVIAR (source) to i-LIDS (target), where three source classes (in red) and four target classes (in blue) are used for demonstration. Different markers indicate different persons (classes). The x-axis and y-axis are the first two PCA scores of the samples in the shared latent space. When there is no CTDD, blue circles and blue hexagrams collapse with red diamonds, blue plus signs collapse with red asterisks. However, after imposing CTDD, data from different tasks are well separated.

samples collapse with the source samples (e.g. blue circles and blue hexagrams collapse with red diamonds, blue plus signs collapse with red asterisks). When imposing CTDD, we obtained the newly learned $\mathbf{W}_0$, and projected the same samples into the shared latent subspace induced by the new $\mathbf{W}_0$ in the same way. As shown in Fig. 3(b), after imposing CTDD, data from different tasks are well separated.

CTDD is essentially an averaged pairwise distance in the shared space between data samples from different tasks. In the case of two tasks (one source task and one target task), the above CTDD model could be simplified as follows:

$$CTDD(\mathbf{W_0}) = \frac{1}{N}\mathrm{tr}(\mathbf{W}_0^{'}\{\sum_{i,j}(\mathbf{x}_i^s - \mathbf{x}_j^t)(\mathbf{x}_i^s - \mathbf{x}_j^t)'\}\mathbf{W}_0) \tag{12}$$

where $\mathbf{x}_i^s$ is the i^{th} sample from source task and $\mathbf{x}_j^t$ is the j^{th} sample from target task. Ideally, we want a large value of cross task data discrepancy.

We incorporate the above CTDD criterion in Eq. (11) into the AMT-DCA developed in the last section and therefore present a new model:

$$\mathbf{W}^* = \arg\max_{\mathbf{w}} \frac{\mathrm{tr}(\mathbf{W}'\mathbf{A}\mathbf{W}) + \alpha CTDD(\mathbf{W}_0)}{\mathrm{tr}(\mathbf{W}'\mathbf{B}\mathbf{W})} \tag{13}$$

where $\mathbf{W}$ follows the same definition in Eq. (9). We call the above model as *constrained asymmetric multi-task discriminant component analysis* (cAMT-DCA). The parameter α indicates the contribution of CTDD in the shared latent space to the overall objective function. By redefining $\mathbf{A} \leftarrow \mathbf{A} + \alpha(\mathbf{\Theta}_0)'\{\frac{1}{N}\sum_{k\neq l}\sum_{i,j}(\mathbf{x}_i^k - \mathbf{x}_j^l)(\mathbf{x}_i^k - \mathbf{x}_j^l)'\}\mathbf{\Theta}_0$, where $\mathbf{\Theta}_0 = [\mathbf{I}_d, \mathbf{O}_{d\times d}, \ldots, \mathbf{O}_{d\times d}] \in \mathbb{R}^{d\times(m+2)d}$, a block matrix where except for the first identity block, all the other blocks are $\mathbf{O}_{d\times d}$ and $m \geq 1$, $\mathbf{W}_0 = \mathbf{\Theta}_0\mathbf{W}$, the solution can be obtained by Eq. (8).

Although the form of CTDD is similar to the scatter matrix of data used in the last section ($S_b^q, S_w^q, q \in \{s,t\}$), the role of CTDD is distinct from theirs. CTDD aims to separate data across tasks (between-task data separation), while the scatter matrices are to compute either the inter-class or intra-class variance within a single task (within-task data separation). On the other hand, the CTDD is a function of $\mathbf{W}_0$ (Eq. (11)) and introduced to constrain the shared latent space $\mathbf{W}_0$ across tasks rather than the entire projection for each task.

5 Experiments

5.1 Datasets and Settings

Datasets. We selected four benchmark datasets in Re-ID: VIPeR [8], 3DPeS [29], i-LIDS [25] and CAVIAR [30]. VIPeR consists of 1264 outdoor images of 632 individuals, with two images of size 128×48 per individual. View angle change is one major cause of appearance change. For instance, most of the matched pairs contain one front/back view and one side-view (see Fig. 4(a)). Brightness change is also present, but there is little occlusion. 3DPeS includes 1011 images of 192 individuals captured on an academic campus, from eight different surveillance cameras with significantly different view angles. Images were collected during different periods of the day, resulting in strong variations of lighting conditions (see Fig. 4(b)). In the i-LIDS dataset, which was captured indoor at a busy airport arrival hall, there are 119 people with a total 476 person images with an average of four images per person. Many of these images undergo large lighting variation, considerable view angle change, and are subject to large occlusions (see Fig. 4(c)). CAVIAR contains 1220 images of 72 individuals captured from two cameras in a shopping center scenario. Images from the second camera present large lighting variation, and blurring effect due to low resolution (see Fig. 4(d)).

Transfer Setting. We set each of these four datasets as *target* dataset. When the target dataset was fixed, we transferred other datasets (called *source* datasets) to each of them separately. Each target dataset was paired with either a single source dataset (termed as *single* transfer) or multiple source datasets (termed as *multiple* transfer). For example, when CAVIAR was used as a target dataset, all the other three datasets (VIPeR, i-LIDS, or 3DPeS) were used as source datasets, either individually or jointly in model learning. In our experiment, we use '$\rightarrow$' to indicate the direction of transfer. For instance, we

(a) VIPeR (b) 3DPeS (c) i-LIDS (d) CAVIAR

Fig. 4: Illustration of person images of the four datasets. Images in the same column are from the same person.

denote "VIPeR→CAVIAR" as the transfer from source VIPeR dataset to target CAVIAR dataset. In total, we have 28 cases of transfer including 12 different cases of single transfer, and 16 different cases of multiple transfer.

Experiment Protocol. For all of the methods in each transfer case, we selected all images in source dataset for training, and randomly split the target dataset into two halves, one half as a *target training* set and the other half as the *target test* set. The performance of matching rate on target test set is the average over 10 random splits. Our split is carried out at the *person* (class) level to ensure that there is no overlap of persons between the target training set and the target test set, i.e., no person participating the training will be seen in the test set.

We randomly selected p images per person in the target training set for training. As a default setting, we set $p = 2$, i.e., only one matched pair per person. This setting is consistent in all of our experiments unless stated otherwise. The purpose of this setting is to test the performance of our algorithm in the case of only limited training samples available in a target task for a Re-ID system. For testing, we adopted a single-shot setting. When each of 3DPeS, i-LIDS and CAVIAR was used as a target dataset, we followed the standard protocol [23]–[26], [36], and randomly selected 1 image from each person in the target test set as a *gallery* image, and the rest as *probe* images. Since VIPeR (and only) has camera view label for each image, so when VIPeR was used as target dataset, for each person in the test set, one image in camera view A was set as gallery image and the other image in camera view B was set as probe image. In this way, the target test set was partitioned into a gallery set and a probe set. The number of gallery images equals to the number of persons in the test set. For each probe image, Re-ID aims to find the best match from the gallery images. We will discuss the performance when more training target images were used in Sec. 5.6.

The performance was evaluated in terms of cumulative matching characteristic curve (CMC), which is a standard measurement for Re-ID [23]–[26], [36]. The CMC curve represents the probability of finding the correct match over the top r in the gallery image ranking, with r varying from 1 to 20.

Feature Representation. In our experiments, all images were resized to 128×48 pixels. We described the appearance of each pedestrian image by a set of three basic features: color, LBP and HOG features, capturing local color, texture and shape respectively. Each type of feature was extracted based on local spatial partition. Specifically, we partitioned each image into a number of overlapping blocks of size 16×16 each, with a step size of every 8 pixels in both the horizontal and vertical directions. We then extracted features from each block. For each block, the color feature was created by concatenating 16-bin histogram of 8 color channels (RGB, YCbCr, HS). HOG features and LBP features were also extracted for each block. Therefore, each block was represented by a 484-dimensional feature vector. For each image, a total of 75 blocks were extracted resulting in a 36300-dimensional feature vector by concatenating all of the block features. These vectors were then compressed into 100-dimensional vectors using PCA before applying all subspace-based methods.

Parameter Setting. For all subspace learning/metric learning methods (including ours) except for RDC [2], we extracted the largest 100 eigenvectors for discriminant modeling, except for the case with VIPeR as target dataset, where we extracted the largest 40 eigenvectors. Setting the final dimension of the learned subspace to 40 performs much better than setting to 100 in this case. As default settings of the proposed cAMT-DCA, we set $\beta = 0.1$, $\gamma = 0.8$. α was set as $1-\beta$, reflecting its role on weighting the common component $\mathbf{W_0}$ in CTDD. This setting is consistent with Eq. (1), and convenient for parameter tuning. Detailed discussions and analysis of these two parameters are presented in Sec. 5.5.

Our methods are compared favorably with the following methods in Re-ID or related, broadly in three categories: 1) six single-task metric learning based methods 2) two multi-task metric learning methods and 3) two domain adaptation methods. All of the methods for comparison used exactly the same features and were conducted under the same training and test setting. Detailed descriptions of the comparison are presented in the subsequent sections.

5.2 cAMT-DCA vs. Single-task Methods

Comparison Protocol. In order to show that whether the learning of *single-task* methods on source datasets could generalize well on target dataset, we selected six recent metric learning algorithms in Re-ID for comparison including LFDA [26], LMNN [17], the popular discriminant distance learning method KISSME [22], LADF [27], PCCA [21], and RDC [25], because our approach is also a subspace method. Note none of them is exclusively designed for transfer. Therefore, for a fair and complete comparison, each of them was trained in three different cases: 1) using target training data only, where the results are shown in Table 1; 2) using source data only for training and then directly applying the learned model to target test set, where the results are shown in Table 2; 3) using a pooled set of source data and target training set for training, where the results are shown in Table 3. In the three different cases, the results of cAMT-DCA were obtained based on a pooled set of source and target training set as required by the algorithm.

It is worth noting that both RDC and PCCA suffer from the huge computational cost with increasing size of training set. In

2. In the RDC model, dimension reduction is not needed. As shown in [25], dimension reduction will degrade the performance drastically.

Methods	VIPeR→i-LIDS				3DPeS→i-LIDS				CAVIAR→i-LIDS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**36.47**	**60.59**	**72.13**	**84.17**	**33.79**	**54.96**	**67.89**	**81.38**	**33.85**	**57.46**	**69.79**	**81.27**
LFDA_T	30.32	51.81	64.46	79.86	30.32	51.81	64.46	79.86	30.32	51.81	64.46	79.86
LMNN_T	27.14	46.61	56.41	74.00	27.14	46.61	56.41	74.00	27.14	46.61	56.41	74.00
KISSME_T	20.31	40.95	53.43	70.11	20.31	40.95	53.43	70.11	20.31	40.95	53.43	70.11
LADF_T	14.20	36.49	49.60	69.59	14.20	36.49	49.60	69.59	14.20	36.49	49.60	69.59
PCCA_T	13.48	34.14	50.30	71.01	13.48	34.14	50.30	71.01	13.48	34.14	50.30	71.01
RDC_T	30.42	51.19	61.88	77.10	30.42	51.19	61.88	77.10	30.42	51.19	61.88	77.10

Methods	VIPeR→CAVIAR				3DPeS→CAVIAR				i-LIDS→CAVIAR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**34.39**	**59.84**	**72.63**	**90.67**	**33.54**	**57.76**	**73.61**	**91.88**	**35.39**	**60.68**	**75.53**	**92.23**
LFDA_T	28.41	49.91	63.79	82.19	28.41	49.91	63.79	82.19	28.41	49.91	63.79	82.19
LMNN_T	24.41	39.71	55.78	79.40	24.41	39.71	55.78	79.40	24.41	39.71	55.78	79.40
KISSME_T	20.28	35.21	52.32	77.18	20.28	35.21	52.32	77.18	20.28	35.21	52.32	77.18
LADF_T	20.68	46.07	62.23	81.55	20.68	46.07	62.23	81.55	20.68	46.07	62.23	81.55
PCCA_T	16.45	37.98	53.81	76.30	16.45	37.98	53.81	76.30	16.45	37.98	53.81	76.30
RDC_T	28.75	45.86	58.55	75.25	28.75	45.86	58.55	75.25	28.75	45.86	58.55	75.25

Methods	VIPeR→3DPeS				i-LIDS→3DPeS				CAVIAR→ 3DPeS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**31.88**	**53.49**	**63.94**	**75.08**	**30.19**	**52.59**	**63.37**	**74.56**	**29.51**	**51.03**	**62.29**	**74.32**
LFDA_T	26.57	48.90	61.42	72.35	26.57	48.90	61.42	72.35	26.57	48.90	61.42	72.35
LMNN_T	23.68	43.91	55.45	67.88	23.68	43.91	55.45	67.88	23.68	43.91	55.45	67.88
KISSME_T	13.96	31.90	44.04	58.68	13.96	31.90	44.04	58.68	13.96	31.90	44.04	58.68
LADF_T	15.53	35.48	49.27	65.28	15.53	35.48	49.27	65.28	15.53	35.48	49.27	65.28
PCCA_T	8.56	25.13	37.55	54.12	8.56	25.13	37.55	54.12	8.56	25.13	37.55	54.12
RDC_T	25.58	44.74	54.59	65.07	25.58	44.74	54.59	65.07	25.58	44.74	54.59	65.07

Methods	i-LIDS → VIPeR				CAVIAR→VIPeR				3DPeS→ VIPeR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**23.39**	**52.75**	**67.12**	**81.14**	**22.18**	**50.44**	**64.94**	**80.32**	**21.61**	**50.92**	**66.27**	**81.36**
LFDA_T	20.89	48.39	63.96	78.51	20.89	48.39	63.96	78.51	20.89	48.39	63.96	78.51
LMNN_T	8.13	21.80	31.52	44.65	8.13	21.80	31.52	44.65	8.13	21.80	31.52	44.65
KISSME_T	20.25	48.01	63.23	79.81	20.25	48.01	63.23	79.81	20.25	48.01	63.23	79.81
LADF_T	9.72	29.53	44.34	61.14	9.72	29.53	44.34	61.14	9.72	29.53	44.34	61.14
PCCA_T	16.65	44.24	61.27	78.45	16.65	44.24	61.27	78.45	16.65	44.24	61.27	78.45
RDC_T	17.78	40.66	52.88	67.18	17.78	40.66	52.88	67.18	17.78	40.66	52.88	67.18

TABLE 1: Matching rate(%): cAMT-DCA vs. single-task methods. '_T' indicates the single-task methods are learned on target datasets only. Two sample images ($p = 2$) are used for each target person.

Methods	VIPeR→i-LIDS				3DPeS→i-LIDS				CAVIAR→i-LIDS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**36.47**	**60.59**	**72.13**	**84.17**	**33.79**	**54.96**	**67.89**	81.38	**33.85**	**57.46**	**69.79**	**81.27**
LFDA_S	31.32	51.93	62.56	79.24	28.42	49.25	62.35	79.58	31.50	53.99	66.71	78.18
LMNN_S	29.16	50.41	63.96	79.19	27.52	46.61	60.32	76.38	29.43	52.37	62.84	76.33
KISSME_S	32.22	51.87	63.34	80.97	27.86	49.46	65.81	**81.65**	30.73	54.23	68.61	80.80
LADF_S	14.16	35.21	49.04	66.44	10.85	34.58	52.99	71.75	9.28	33.66	46.35	64.14
PCCA_S	22.83	40.97	54.41	71.25	23.55	46.44	61.45	80.02	19.64	43.20	59.31	76.77

Methods	VIPeR→CAVIAR				3DPeS→CAVIAR				i-LIDS→CAVIAR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**34.39**	**59.84**	**72.63**	**90.67**	**33.54**	**57.76**	**73.61**	**91.88**	**35.39**	**60.68**	**75.53**	**92.23**
LFDA_S	32.43	51.82	64.73	83.66	30.09	52.70	67.94	84.80	33.91	53.14	67.02	87.38
LMNN_S	28.01	48.40	64.56	84.16	27.18	47.59	63.04	83.57	28.97	48.09	64.04	84.04
KISSME_S	30.19	52.45	67.62	84.37	30.60	52.81	67.86	84.03	30.69	53.58	70.26	88.08
LADF_S	25.08	50.17	65.04	82.02	18.65	46.27	60.33	83.46	25.48	51.52	67.65	84.13
PCCA_S	23.07	41.67	57.47	83.27	24.04	46.79	61.91	83.51	20.78	50.12	69.50	85.64

Methods	VIPeR→3DPeS				i-LIDS→3DPeS				CAVIAR→ 3DPeS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**31.88**	**53.49**	**63.94**	**75.08**	**30.19**	**52.59**	**63.37**	**74.56**	**29.51**	**51.03**	**62.29**	**74.32**
LFDA_S	26.85	46.18	55.88	66.36	25.41	43.75	53.66	65.30	26.48	45.49	54.50	65.32
LMNN_S	26.93	47.04	56.12	66.72	24.43	43.20	52.04	63.00	25.72	44.57	53.94	64.74
KISSME_S	27.64	47.48	56.14	67.28	25.74	45.60	56.35	68.36	26.91	46.33	55.52	66.24
LADF_S	12.23	32.28	43.32	57.83	11.85	28.90	41.05	56.51	6.49	17.84	27.33	42.63
PCCA_S	19.67	39.70	51.11	63.93	17.03	35.72	47.90	63.09	16.53	35.31	46.30	61.86

Methods	i-LIDS → VIPeR				CAVIAR→VIPeR				3DPeS→ VIPeR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**23.39**	**52.75**	**67.12**	**81.14**	**22.18**	**50.44**	**64.94**	**80.32**	**21.61**	**50.92**	**66.27**	**81.36**
LFDA_S	8.16	22.47	33.32	44.59	8.23	21.11	30.06	43.26	8.64	22.18	33.61	48.10
LMNN_S	7.06	23.01	34.59	46.30	7.63	20.82	31.20	44.97	6.46	18.23	27.85	40.38
KISSME_S	8.13	22.15	31.96	44.78	9.87	20.00	29.37	41.65	6.87	20.95	29.43	42.94
LADF_S	2.72	10.35	17.85	28.35	1.08	4.94	9.91	16.71	3.04	11.11	19.68	31.68
PCCA_S	5.57	16.58	23.26	33.39	5.57	13.29	20.57	31.23	5.54	16.77	26.71	39.18

TABLE 2: Matching rate(%): cAMT-DCA vs. single-task methods. '_S' indicates the single-task methods are learned on source datasets only. Two sample images ($p = 2$) are used for each target person.

Methods	VIPeR→i-LIDS				3DPeS→i-LIDS				CAVIAR→i-LIDS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**36.47**	**60.59**	**72.13**	**84.17**	**33.79**	**54.96**	**67.89**	**81.38**	**33.85**	**57.46**	**69.79**	**81.27**
LFDA-Mix	31.82	51.59	63.96	80.24	30.10	51.26	63.30	78.86	30.53	49.62	62.39	79.03
LMNN-Mix	30.15	51.20	63.57	79.98	27.69	47.84	60.33	75.95	29.26	49.30	62.23	76.17
KISSME-Mix	35.24	54.95	67.54	83.32	26.87	45.22	58.38	75.17	27.35	44.65	57.27	73.60
LADF-Mix	16.18	38.51	52.00	69.85	11.67	38.72	57.41	76.09	14.55	38.12	52.56	68.60
PCCA-Mix	23.96	47.39	62.06	77.85	18.02	44.51	61.40	78.92	20.04	45.74	59.78	74.94

Methods	VIPeR→CAVIAR				3DPeS→CAVIAR				i-LIDS→CAVIAR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**34.39**	**59.84**	**72.63**	**90.67**	**33.54**	**57.76**	**73.61**	**91.88**	**35.39**	**60.68**	**75.53**	**92.23**
LFDA-Mix	32.32	53.39	65.44	85.22	31.12	50.99	65.60	85.64	33.70	53.66	69.41	87.56
LMNN-Mix	27.80	49.62	65.00	85.17	27.05	46.87	62.15	83.45	27.94	47.20	62.07	82.55
KISSME-Mix	32.11	53.30	67.96	85.89	27.64	45.61	60.50	81.59	30.76	50.89	67.51	86.65
LADF-Mix	25.85	50.85	66.59	84.38	25.85	50.85	66.59	84.38	30.41	56.04	70.28	88.67
PCCA-Mix	25.63	48.43	64.26	85.79	24.72	49.69	67.73	87.64	26.38	52.26	69.20	88.01

Methods	VIPeR→3DPeS				i-LIDS→3DPeS				CAVIAR→ 3DPeS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**31.88**	**53.49**	**63.94**	**75.08**	**30.19**	**52.59**	**63.37**	**74.56**	**29.51**	**51.03**	**62.29**	**74.32**
LFDA-Mix	27.38	48.48	58.79	69.59	26.82	48.85	60.21	71.79	23.79	43.43	54.59	66.57
LMNN-Mix	27.44	47.92	58.01	69.42	24.92	45.64	55.59	67.28	25.29	45.15	55.62	67.75
KISSME-Mix	28.94	49.82	60.66	71.28	26.31	47.00	59.51	71.50	22.34	39.81	51.20	63.26
LADF-Mix	13.13	34.15	47.76	63.35	9.25	27.55	41.86	59.53	10.29	26.32	39.80	54.82
PCCA-Mix	22.39	45.66	58.18	71.89	22.36	44.23	56.63	71.97	19.32	40.26	52.38	67.84

Methods	i-LIDS → VIPeR				CAVIAR→VIPeR				3DPeS→ VIPeR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**23.39**	**52.75**	**67.12**	**81.14**	**22.18**	**50.44**	**64.94**	**80.32**	**21.61**	**50.92**	**66.27**	**81.36**
LFDA-Mix	19.24	45.44	59.18	75.25	16.68	40.73	56.61	72.47	16.90	42.63	58.23	74.78
LMNN-Mix	8.13	21.93	33.45	46.17	7.72	21.17	31.36	46.27	7.78	20.89	31.30	44.46
KISSME-Mix	15.03	35.47	49.34	64.46	9.05	20.66	29.72	39.94	12.22	32.18	44.15	58.48
LADF-Mix	6.61	21.11	33.58	49.08	6.30	21.42	33.45	49.15	8.96	28.70	42.85	58.83
PCCA-Mix	14.34	41.61	56.71	72.37	-	-	-	-	14.37	39.94	55.60	72.34

TABLE 3: Matching rate(%): cAMT-DCA vs. single-task methods. '-Mix' indicates the single-task methods are learned on a pooled set of source and target datasets. Two sample images ($p = 2$) are used for each target person.

Methods	VIPeR→i-LIDS				3DPeS→i-LIDS				CAVIAR→i-LIDS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**36.47**	**60.59**	**72.13**	84.17	**33.79**	54.96	67.89	**81.38**	**33.85**	**57.46**	**69.79**	**81.27**
AMT-DCA	35.75	57.85	70.51	**84.39**	32.84	**55.13**	**68.11**	80.53	32.55	53.89	66.48	79.80

Methods	VIPeR→CAVIAR				3DPeS→CAVIAR				i-LIDS→CAVIAR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**34.39**	**59.84**	**72.63**	**90.67**	**33.54**	**57.76**	**73.61**	**91.88**	**35.39**	**60.68**	**75.53**	92.23
AMT-DCA	33.45	55.42	70.81	89.52	33.14	56.18	71.14	91.72	33.87	58.75	73.35	**92.52**

Methods	VIPeR→3DPeS				i-LIDS→3DPeS				CAVIAR→ 3DPeS			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**31.88**	**53.49**	**63.94**	**75.08**	**30.19**	**52.59**	**63.37**	**74.56**	**29.51**	**51.03**	**62.29**	**74.32**
AMT-DCA	30.48	52.45	62.49	73.72	29.43	51.23	62.63	73.66	27.59	48.26	59.16	71.08

Methods	i-LIDS→VIPeR				CAVIAR→VIPeR				3DPeS→ VIPeR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**23.39**	**52.75**	**67.12**	**81.14**	**22.18**	**50.44**	**64.94**	**80.32**	**21.61**	**50.92**	**66.27**	**81.36**
AMT-DCA	21.36	50.54	66.20	81.08	20.35	48.13	62.94	77.12	20.13	49.68	65.25	79.11

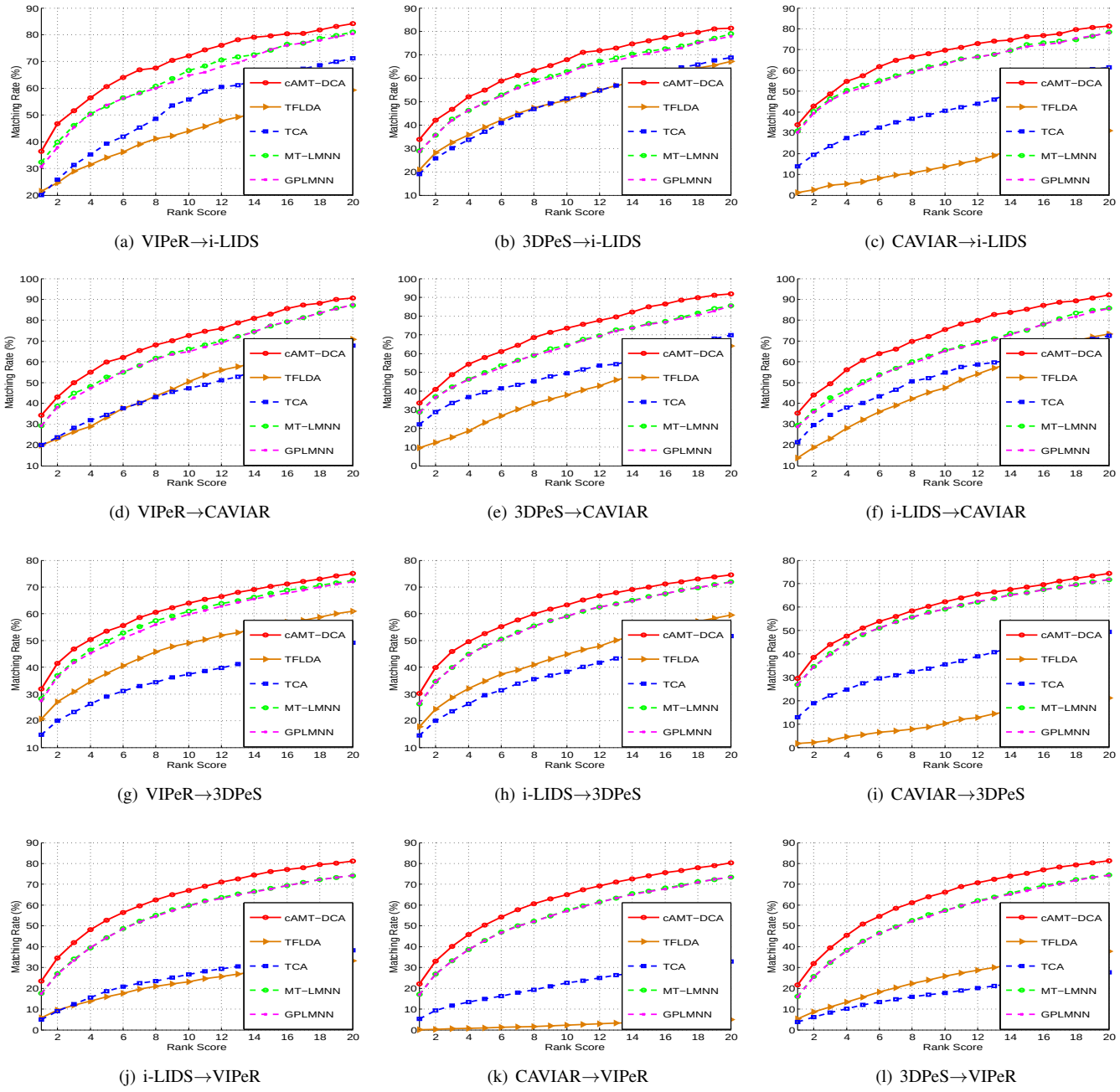
TABLE 4: Matching rate(%): With and Without CTDD in cAMT-DCA. The AMT-DCA is exactly cAMT-DCA without using CTDD. Two sample images ($p = 2$) are used for each target person.

all multiple transfer cases, RDC-Mix and PCCA-Mix could not be successfully tested on the pooled source training data and target training data, even on a high performance computing platform with an Intel-16-core CPU and 144GB RAM. We have observed that PCCA-Mix costs 120GB RAM on average and usually gets the server stalled. The computational complexity of RDC is even much higher than PCCA. Therefore, we tested RDC trained on target dataset only. In Table 3, in "CAVIAR→VIPeR", PCCA-Mix could not be tested as well due to high computational cost.

For PCCA, we tuned the parameter β in the generalized logistic loss function [21] in a wide range and reported the best results at $\beta = 3$.

Performance Evaluation. In all transfer settings, cAMT-DCA performs notably better than all of them, and improves a lot at $r = 1$, with about 7% increase over LFDA_T on "i-LIDS→CAVIAR" and about 6% increase on "VIPeR→i-LIDS", as shown in Table 1. In Table 3, cAMT-DCA achieves about 5% improvement at $r = 1$ over LFDA-Mix on "VIPeR→i-LIDS", "CAVIAR→VIPeR" and "3DPeS→VIPeR", and about 6% improvement on " CAVIAR→3DPeS". As r increases, the matching rate of all methods increases and cAMT-DCA consistently outperforms others, with approximately 10% improvement over LFDA_T on "VIPeR→CAVIAR", over LFDA_S on "VIPeR→i-LIDS", and over LFDA-Mix on "VIPeR→i-LIDS" at $r = 5$ and $r = 10$, as shown in Table 1, Table 2 and Table 3.

When VIPeR was used as a target dataset, compared with the other transfer cases, the performances of all methods are overall lower. One of the reasons is that there are 316 persons in the test set, in comparison with an average of 60 persons in the test set for each of other datasets. The larger the number of persons in the test set is, the harder for the probe image to find the correct match from the gallery images.



(a) VIPeR→i-LIDS (b) 3DPeS→i-LIDS (c) CAVIAR→i-LIDS

(d) VIPeR→CAVIAR (e) 3DPeS→CAVIAR (f) i-LIDS→CAVIAR

(g) VIPeR→3DPeS (h) i-LIDS→3DPeS (i) CAVIAR→3DPeS

(j) i-LIDS→VIPeR (k) CAVIAR→VIPeR (l) 3DPeS→VIPeR

Fig. 5: Matching rates of cAMT-DCA, multi-task methods and domain adaptation methods, with i-LIDS,CAVIAR, 3DPeS and VIPeR as target dataset. Two sample images ($p = 2$) are used for each target person.

Discussion. Except for the case with VIPeR as target dataset, it is observed that only using the source dataset for the chosen metric learning algorithms often results in better performance than only using limited target data in other cases. It is probably due to the following reasons.

Firstly, the number of training images per person would drastically affect the modeling of metric learning algorithm when the size of the target training set is very small. In this case there are not enough pairs of intra-class samples for the estimation of various intra-person variations for cross-view person images. On the contrary, the number of intra-person pairs in source dataset is much larger than that of target dataset. For example, the number of intra-person pairs is 866, 10490, 3475, and 632 when each of i-LIDS, CAVIAR, 3DPeS and VIPeR is used as source dataset, in contrast to 59, 36, 96 and 316 when they are used as target datasets, respectively. In those cases, when source datasets could provide more samples for each person, it might lead to a more reliable estimation of the parameters in metric learning methods.

Second, although source and target datasets are from different scenarios, they are not completely irrelevant. Instead they all present the variations of one big category of "person". Intuitively, the knowledge (e.g. feature projections) learned to recognize and differentiate persons in one group of people could be used to differentiate persons in another group.

The above could be the potential reasons those non-transfer

(a) CAVIAR→i-LIDS (b) VIPeR→3DPeS

Fig. 6: Sample results of Person Re-ID over the same probe image using cAMT-DCA (top row), LFDA-Mix (middle row) and MT-LMNN (bottom row). In each row, the left-most is the probe image; images in the middle are the top 10 matched gallery images, with a red box highlighting the correct match, and the right-most shows the ground truth.

Methods	$p=3$				$p=4$				$p=5$			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**36.63**	**58.72**	**74.12**	**89.94**	**37.09**	59.90	**75.34**	91.39	**39.22**	**61.88**	**76.55**	89.88
LFDA-Mix	32.73	53.61	66.82	84.78	33.72	54.47	68.03	86.19	34.03	55.10	68.39	86.57
LMNN-Mix	28.34	47.60	62.25	82.87	29.99	47.98	61.53	83.05	28.63	46.81	61.57	83.21
KISSME-Mix	32.89	58.58	72.13	88.02	33.97	**60.62**	74.51	**92.64**	36.91	60.84	74.92	**91.18**
LADF-Mix	23.46	51.65	67.81	83.68	20.76	49.90	67.01	87.26	26.20	56.31	72.38	88.33
PCCA-Mix	27.68	53.52	68.11	87.64	25.81	54.86	72.25	90.15	27.93	55.61	71.37	89.36
TCA	19.10	37.13	50.84	73.94	19.68	36.27	49.59	77.96	19.05	36.58	51.70	74.40
TFLDA	18.67	33.43	49.09	70.68	20.05	33.40	49.27	70.94	19.81	33.42	48.76	71.16
MT-LMNN	29.85	52.90	68.40	85.64	30.92	49.71	64.11	84.55	29.00	51.05	66.22	85.90
GPLMNN	30.04	52.98	68.58	86.37	30.16	49.51	63.63	84.91	29.52	49.83	64.93	85.86

TABLE 5: cAMT-DCA vs. others: matching rate(%) in "VIPeR→CAVIAR", with respect to different number p of target training images for each person.

Methods	VIPeR+CAVIAR→i-LIDS				VIPeR+3DPeS→CAVIAR				VIPeR+i-LIDS→3DPeS				CAVIAR+i-LIDS→VIPeR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**35.64**	**58.86**	**70.45**	**83.72**	**33.70**	**58.20**	**75.68**	**93.45**	**31.86**	**52.37**	**63.06**	**73.29**	**20.35**	**48.26**	**63.01**	**77.94**
LFDA-Mix	30.27	51.20	64.57	80.07	32.14	53.10	64.73	85.28	28.00	49.50	58.68	70.15	17.25	42.94	58.20	73.13
LMNN-Mix	30.15	51.09	63.57	77.34	26.99	47.49	61.60	83.66	25.93	45.41	54.99	67.05	8.32	21.14	32.44	46.36
KISSME-Mix	26.58	48.57	59.94	75.89	30.58	49.76	61.84	83.94	27.64	48.58	58.65	70.07	8.77	21.68	31.08	41.84
LADF-Mix	18.86	42.35	55.90	71.46	22.76	51.48	68.80	90.11	9.40	27.93	40.79	58.88	6.80	20.76	32.15	47.63
MT-LMNN	31.39	54.44	66.88	81.48	29.14	51.75	65.87	88.42	28.43	49.12	60.19	71.49	16.33	43.61	58.04	72.56
GPLMNN	32.00	52.98	65.70	80.98	29.47	50.70	63.47	87.69	27.26	48.20	59.05	70.75	16.20	43.07	57.37	72.72
TCA	14.51	32.70	44.65	64.55	21.87	41.25	53.79	74.61	14.84	27.85	37.89	50.22	5.89	16.55	24.43	37.06
TFLDA	21.04	41.79	52.88	68.72	18.54	34.71	49.04	73.45	18.28	35.57	46.10	57.66	5.16	13.54	19.68	29.21

TABLE 6: cAMT-DCA vs. others: matching rate(%) with i-LIDS, CAVIAR, 3DPeS and VIPeR as target dataset each, and two of others are used as sources for transfer. Two sample images ($p=2$) are used for each target person.

methods trained only using source datasets often perform better than using target only, when i-LIDS, 3DPeS or CAVIAR was used as target.

However, simply training a method on source only is not the best way to tackle the problem of the cross-scenario person re-identification. It is because these estimations on source are not ideal for the target task without identifying the latent features shared between source and target sets. For these non-transfer methods, what they learned on source dataset is not all what are desired on target. This can be evidenced by the lower matching rate with less than 10% at rank 1 when directly applying these non-transfer models learned on source datasets to target dataset VIPeR as shown in Table 2.

In addition, it could be observed that LFDA-Mix, KISSME-Mix, LADF-Mix and LMNN-Mix perform almost the same as LFDA_S, KISSME_S, LADF-S and LMNN_S do in some cases. Sometimes, they perform even worse on '3DPeS→i-LIDS", "CAVIAR→i-LIDS", and "3DPeS→CAVIAR" (see Table 2 and Table 3). It shows that simply pooling all data from different tasks together would not always help improve the Re-ID performance on target. This further indicates that cross-scenario differences indeed exist, and existing single-task methods did not specifically consider the essential discrepancy across tasks. Therefore they are largely biased by source datasets. On the contrary, our proposed cAMT-DCA can efficiently identify the shared latent features and get it transferred to target task.

5.3 cAMT-DCA vs. Multi-task + Domain Adaptation Methods

Our proposed method is an asymmetric multi-task learning, so here we compared some representative multi-task learning methods, which are subspace/metric based. Multi-Task LMNN (MT-LMNN) [33] and Geometry Preserving LMNN (GPLMNN) [34] were selected. The regularization parameter for the common metric and task-specific metric was set to 1 for MT-LMNN, 0.5 for GPLMNN, achieving their best performance. Von Neumann divergence was used in GPLMNN.

Methods	VIPeR+CAVIAR+3DPeS→i-LIDS				VIPeR+i-LIDS+3DPeS→CAVIAR				VIPeR+CAVIAR+i-LIDS→3DPeS				CAVIAR+i-LIDS+3DPeS→VIPeR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**36.53**	**59.31**	**70.50**	**84.39**	**36.22**	**59.82**	**75.16**	**92.58**	**30.52**	**51.68**	**61.43**	**72.28**	**19.49**	**46.77**	**61.87**	**77.72**
LFDA -Mix	32.51	53.43	66.97	81.37	33.07	53.95	66.49	85.43	26.04	45.18	56.76	68.78	16.80	42.56	56.93	72.28
LMNN-Mix	31.21	50.69	63.01	77.51	28.30	48.30	62.57	83.37	26.22	45.18	55.25	66.94	8.45	22.34	32.15	46.27
KISSME-Mix	29.21	48.06	63.40	78.13	31.85	51.76	66.49	85.30	25.96	44.16	53.79	66.62	9.87	22.18	31.87	44.72
LADF-Mix	16.94	42.97	57.86	72.86	22.27	52.31	71.17	89.36	13.45	34.43	47.57	63.78	6.65	19.08	30.54	44.84
MT-LMNN	31.78	55.56	66.60	81.37	30.58	52.92	67.54	87.22	29.14	49.70	60.57	70.97	16.36	42.28	56.20	72.31
GPLMNN	32.79	53.26	64.86	81.42	30.16	50.81	65.15	87.82	27.89	48.72	59.19	70.94	16.30	41.84	55.57	71.96
TCA	16.20	35.01	46.50	65.32	20.75	38.16	55.02	78.78	14.61	28.49	37.69	50.64	4.56	14.11	21.65	30.89
TFLDA	24.41	41.53	55.98	70.58	19.03	34.21	49.88	73.04	19.30	34.09	43.75	56.12	4.91	13.20	21.58	31.71

TABLE 7: cAMT-DCA vs. others: matching rate(%) with i-LIDS, CAVIAR, 3DPeS and VIPeR as target dataset each, and the other three are used as sources for transfer. Two sample images ($p=2$) are used for each target person.

Methods	VIPeR→CAVIAR				i-LIDS→CAVIAR				3DPeS→CAVIAR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**65.97**	**87.15**	**93.66**	**98.97**	**64.00**	**83.31**	**93.75**	98.31	**65.93**	**84.26**	**92.55**	**98.60**
LFDA-Mix	62.46	85.55	92.22	98.42	61.70	82.31	90.99	**98.69**	62.29	84.04	91.80	97.59
LMNN-Mix	58.58	81.78	90.45	97.06	56.65	80.05	89.52	97.02	57.11	80.20	88.97	97.11
KISSME-Mix	63.58	84.87	92.26	98.20	59.02	82.46	91.51	97.06	58.23	80.68	90.18	96.78
LADF-Mix	34.05	70.54	85.09	97.11	39.32	71.02	84.20	95.48	31.80	64.75	80.51	93.61
PCCA-Mix	46.80	75.02	86.60	95.91	43.94	73.30	84.91	96.44	51.53	77.19	88.83	96.98
MT-LMNN	61.56	83.05	91.58	97.41	58.45	81.54	90.38	97.19	58.75	81.25	90.79	97.15
GPLMNN	60.12	82.70	90.51	97.19	58.01	81.25	89.87	97.22	58.67	80.84	91.03	97.41
TCA	45.08	67.12	78.34	91.11	46.70	69.93	80.35	90.17	45.44	70.65	81.58	92.63
TFLDA	42.57	64.31	77.27	88.66	27.81	50.77	64.61	82.07	48.81	68.57	81.55	93.10

TABLE 8: Matching rate(%) in a multishot setting, single transfer with CAVIAR as target dataset. Two sample images ($p=2$) are used for each person in target training set. Five images are chosen as gallery images for each person in the target test set.

Albeit different from domain adaptation, in order to show typical domain adaptation methods do not work well for cross-scenario person re-identification, we compared two representative subspace-based domain adaptation methods: Transfer Component Analysis (TCA) [37] and Transfer Fisher Linear Discriminant Analysis (TFLDA) [38]. TCA is an unsupervised method without utilizing discriminant information. TFLDA considers differentiating different classes from the source dataset in a subspace, but simultaneously minimizing the Bregman divergence between two distributions of data from the source and target dataset. The regularization parameter γ in TFLDA was carefully tuned, the best result was reported when $\gamma = 0.5$. The comparison results are shown in Fig. 5. It can be seen that, cAMT-DCA obtains the highest matching rate, with an overall performance gain of $10\%-20\%$ over TCA and TFLDA and up to 10% over MT-LMNN and GPLMNN.

These results indicate that 1) the domain adaptation based models are not suitable for the cross-scenario transfer in our case, since the people identities (classes) of the source and target datasets are usually different in cross-scenario transfer Re-ID, while the domain adaptation models are designed to minimize the distribution bias of samples from two domains for the same group of people; 2) the proposed asymmetric multi-task learning is more effective for Re-ID. In addition, unlike other multi-task models, the proposed cAMT-DCA is not iterative and further exploits the discriminant information across tasks in the shared latent space with CTDD constraint enhancing inter-class variation there.

To further illustrate the advantage of the proposed cAMT-DCA, we show some real matching examples of our method, LFDA-Mix, and MT-LMNN over the same probe image, on "CAVIAR→i-LIDS" (see Fig. 6(a)), "VIPeR→3DPeS" (see Fig. 6(b)) [3]. Our proposed cAMT-DCA ranks a correct match higher (i.e., much closer to the probe image) than LFDA-Mix and MT-LMNN. [4]

5.4 With vs. without CTDD in cAMT-DCA

Table 4 shows the results of the proposed model with and without CTDD, where "AMT-DCA" indicates the proposed asymmetric multi-task discriminant component analysis without CTDD. We performed a paired sample z test on the results on Table 4 and confirmed that the improvement is statically significant with computed p-value approximately zero. In particular, the improvement with CTDD over without CTDD is around 3% in many cases, especially when VIPeR and CAVIAR were used as the source datasets. Note that images in VIPeR dataset have large intra-class variations mainly due to viewpoint changes. This implies that the distribution of person images of the source dataset is likely to overlap with that of the target data in the shared latent space. The purpose of CTDD is to reduce such a kind of overlap in the latent subspace. From Fig. 3, we can see that after using CTDD, the overlap between the distributions of images from source dataset and target training set has been noticeably reduced. And the results in Table 4 confirm the efficacy of the idea of CTDD.

5.5 Effect of Parameters

In the proposed model, we introduced two important parameters β and γ to control the coupling of source and target datasets. To test the effect of those parameters, we evaluated cAMT-DCA by varying β in the range of $[0 : 0.1 : 1]$ and γ in the range of $[0.1 : 0.1 : 1]$ [5]. To best visualize the performances, we use AUC (Area Under the CMC Curve)

3. Due to space limitation, the matching example on "VIPeR→CAVIAR" is shown in Fig. 1 in the supplementary file.

4. We also present the Re-ID results of a sample set of probe images using cAMT-DCA on "CAVIAR→i-LIDS", "VIPeR→CAVIAR", and "VIPeR→3DPeS", in Fig. 2, Fig. 3 and Fig. 4, respectively in the supplementary file.

5. The case when $\gamma = 0$ is excluded and not applicable to our method, as $\gamma = 0$ indicates that target dataset is not used in our modeling.

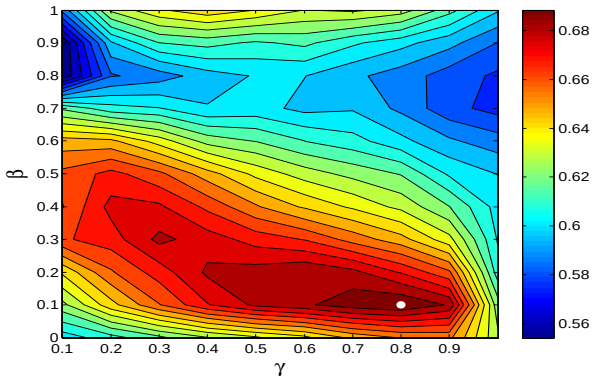


Fig. 7: Visualization of AUC contour parameterized by β and γ in VIPeR→i-LIDS, the highest AUC value is highlighted by the white spot in the figure. Two sample images are used for each target person.

to quantify the performance [6] , which was shown using CMC before. We plotted contours to visualize the change of AUC values with different combinations of β and γ, Fig. 7 shows the AUC contour plot in 'VIPeR→i-LIDS'. The highest AUC value (marked at a white spot when $\beta = 0.1$ and $\gamma = 0.8$) is located in the region with strongest redness.

We have tested the effect of these two parameters in all other transfer cases, and similar conclusions can be drawn, which indicates that the parameters setting is quite reliable across different datasets. Therefore, $\beta = 0.1, \gamma = 0.8$ were set consistently for the experiments in all transfer cases.

5.6 Effect of Number of Target Training Samples

In this experiment, we tested the performance of our algorithm by varying the number of training images per person from 3 to 5 in the *target* training set. It is noted that, different persons have different numbers of instance images in those datasets, e.g., varying from 2 to 8 in the i-LIDS dataset. For those persons in the target training set, if the maximum number of instance images in a person category is less than p (p = 3, 4, 5), all of the images in that person category were used for training. The testing protocol is the same as that of previous experiments.

Here for the clarity of presentation, we report the matching rates against different numbers of target training samples p per person on "VIPeR→CAVIAR" in Table 5[7]. In most of the cases, the proposed method outperforms other methods using exactly the same training set. This further shows that the proposed cAMT-DCA could still be a preferable choice when more training samples are available in the target dataset.

5.7 Transfer from Multiple Source Datasets

We report the results when two source datasets were available in Table 6 and when three source datasets were available in Table 7. Compared with MT-LMNN and GPLMNN, cAMT-DCA achieves an improvement of $3\% - 4\%$ when $r = 1$. In particular, cAMT-DCA performs significantly better than TCA and TFLDA, with a gain up to 20% at $r = 1$. This further shows that the proposed asymmetric multi-task model is more suitable for cross scenario Re-ID than the related ones.

Through the experiments, it is observed that using more source datasets does not necessarily mean a better improvement. For example, as shown in Table 1, 6 and 7, always using VIPeR as source is better than using more except for the case "VIPeR + CAVIAR + 3DPeS → i-LIDS" and " VIPeR + i-LIDS + 3DPeS → CAVIAR". It is probably due to the diversity of source scenarios, when using more source tasks to transfer, the sharing among them may become more ambiguous, and the transfer learning task becomes more challenging.

However, there may still exist room for developping algorithms to select a suitable source dataset or a set of sources for transfer learning for better performance. It is closely related to the theoretical bottleneck about task selection in machine learning. Indeed determining the optimal number of tasks for multi-task modeling remains an open problem in both machine learning and computer vision fields. A future research breakthrough in theory might solve this problem.

5.8 cAMT-DCA in Multi-shot Setting

Thus far, all of our experiments were conducted in a single-shot setting. To further evaluate the proposed method in a multi-shot setting, we compared it with the related approaches in the transfer cases where CAVIAR was used as target dataset (CAVIAR is the best dataset for conducting the experiment under the multi-shot setting as each person has a minimum of 10 images). In the test set, we randomly chose **five** images per person as *gallery* images, and the other images were set as *probe* images. All the other settings are the same as the presented in Sec. 5.1. The results are shown in Table 8 and Table 9.

Methods	VIPeR+3DPeS→CAVIAR				VIPeR+i-LIDS+3DPeS→CAVIAR			
	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
cAMT-DCA	**64.61**	**84.06**	**92.52**	**99.45**	**64.46**	**85.11**	**92.90**	**99.03**
LFDA-Mix	62.81	82.90	91.16	98.18	62.44	83.71	90.53	98.20
LMNN-Mix	57.70	80.66	89.83	97.19	58.08	81.76	91.08	96.97
KISSME-Mix	60.00	80.48	89.54	97.22	63.67	83.47	91.12	96.86
LADF-Mix	39.50	66.83	83.13	93.62	39.64	66.09	82.03	94.52
MT-LMNN	60.14	83.21	92.09	97.94	60.60	82.81	91.89	98.58
GPLMNN	59.57	83.34	91.34	97.63	59.52	82.22	91.78	97.65
TCA	44.58	70.09	82.00	93.94	42.64	67.88	80.16	92.27
TFLDA	45.08	65.41	78.17	89.26	45.07	65.01	77.33	90.40

TABLE 9: Matching rate(%) in a multishot setting, multiple transfer with CAVIAR as target dataset. Two sample images ($p = 2$) are used for each person in target training set. Five images are chosen as gallery images for each person in the target test set.

As expected, all methods have gained notable improvement in the multi-shot setting compared to the corresponding results in the single-shot setting (see Table 3, 6, 7, and Fig. 5). And our proposed cAMT-DCA still outperforms all of them.

6 Conclusion

We have addressed the problem of person re-identification across different scenarios and proposed a constrained asym-

6. The larger the area under CMC curve (AUC) is, the more steep when the rank is small. Hence, a larger AUC always implies a better matching performance over all rank matching.

7. Due to space limitation, the results in "CAVIAR→i-LIDS", and " VIPeR →3DPeS" are presented in the supplementary file.

metric multi-task discriminant component analysis (cAMT-DCA) for leveraging source datasets (i.e. existing datasets) to improve performance of the target task. In particular, in cAMT-DCA, we have explored a cross-task data discrepancy (CTDD) constraint in order to learn a discriminant shared component across tasks that reduces the overlap between the cross-task (inter-class) data. Extensive results have shown the proposed asymmetric multi-task learning approach outperforms seven different state-of-the-art approaches in the case of cross-scenario transfer person re-identification. Our study shows that with limited training samples in target task, it is possible to build up an efficient target Re-ID system in a new surveillance system by a cross-scenario transfer modeling, helping avoid the need to re-collect a lot of labeled data for training.

It is worth noting that an interesting but very challenging problem is unveiled by our study: how to select the most important source datasets to transfer, which is similar to the largely unsolved problem in existing multi-task modeling that how many tasks are enough and which of them is useful.

ACKNOWLEDGMENTS

This project was supported by Natural Science Foundation Of China (No. 61102111, 61472456), and in part by RSE-NSFC joint project (RSE Reference: 443570/NNS/INT).

REFERENCES

[1] P. Agrawal and P. Narayanan, "Person de-identification in videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 299–310, 2011.

[2] S. J. Krotosky and M. M. Trivedi, "Person surveillance using visual and infrared imagery," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1096–1105, 2008.

[3] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 26–33.

[4] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *Eur. Conf. Computer Vision*, 2006, pp. 125–136.

[5] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *Proc. Int. Conf. Pattern Recognition*, vol. 3, 2006, pp. 1204–1207.

[6] P. Dollár, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[7] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE Int. Conf. Computer Vision*, 2007, pp. 1–8.

[8] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Computer Vision*, 2008, pp. 262–275.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[10] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. Eur. Conf. Computer Vision, Workshops and Demonstrations*, 2012, pp. 391–401.

[11] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proc. Eur. Conf. Computer Vision, Workshops and Demonstrations*, 2012, pp. 413–422.

[12] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[13] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

[14] ——, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 2528–2535.

[15] ——, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[16] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[17] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Advances in Neural Information Processing Systems*, 2005, pp. 1473–1480.

[18] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Machine Learning*, 2007, pp. 209–216.

[19] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. British Machine Vision Conf.*, 2010, pp. 21.1–21.11.

[20] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. Asia Conf. Computer Vision*, 2011, pp. 501–512.

[21] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672.

[22] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[23] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. Eur. Conf. Computer Vision*, 2012, pp. 780–793.

[24] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.

[25] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[26] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[27] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[28] R. Caruana, *Multitask Learning*. Springer, 1998.

[29] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Proc. joint ACM Workshop Human Gesture and Behavior Understanding*, 2011, pp. 59–64.

[30] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. British Machine Vision Conf.*

[31] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Processing*, vol. 23, no. 8, 2014.

[32] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Computer Vision*, 2014, pp. 1–16.

[33] S. Parameswaran and K. Q. Weinberger, "Large margin multi-task metric learning," in *Proc. Advances in Neural Information Processing Systems*, 2010, pp. 1867–1875.

[34] P. Yang, K. Huang, and C.-L. Liu, "Geometry preserving multi-task metric learning," *Machine Learning*, vol. 92, no. 1, pp. 133–175, 2013.

[35] W.-S. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: from person to set-based verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2650–2657.

[36] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asia Conf. Computer Vision*, 2012, pp. 31–44.

[37] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[38] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 7, pp. 929–942, 2010.

[39] B. Geng, D. Tao, and C. Xu, "Daml: Domain adaptation metric learning."

[40] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction, 2006," in *Proc. Int. Conf. Machine Learning*, 2006, pp. 905–912.